OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Guest Editorial

Data Management for Integrative Biology

THIS FIRST ISSUE OF VOLUME 7 OF *OMICS A Journal of Integrative Biology* is devoted to the special topic of Data Management. The issue arose from the Workshop on Data Management for Molecular and Cell Biology held at the National Library of Medicine on February 2–3, 2003. The purpose of the workshop was to formulate a research agenda for the data management community to develop better technology for supporting bioinformatics applications. This present issue of the Journal contains both a summary of the workshop report and a collection of the white papers submitted by attendees of the workshop.

The impetus for this workshop was the increased demand for data management systems occasioned by the industrialization of molecular and cell biology over the past 15 years. The development of high throughput technologies for sample preparation, sequencing, microarrays, proteomics, and combinatorial chemistry has led to an explosion in the amount and types of data available for biomedical research. Metabolic models provide a means for linking genomics to pharmacology. It is widely anticipated that these technologies will find clinical applications within a few years, leading to further increases in data volumes. Data management systems are needed to manage such large datasets. Without suitable tools for storing and querying these large datasets, many of the benefits of these massive investments in data collection will be delayed or squandered. Consider the limited utility of the human genome sequence if our only access to it were by reference to printed copies, rather than online approximate sequence matching.

Bioinformaticists have been largely dependent on hand-me-down relational database technology from the business sector. But bioinformatics applications have distinct data management requirements: a wide diversity of data types (sequences, graphs, 3D structures, etc.), extensive use of similarity and pattern matching queries, and a need for data provenance tracking. Furthermore, there is a need to support large scale (e.g., 500 databases) data integration, the associated terminology management, and rapid schema evolution. The data integration problems of integrating large numbers of databases cannot be met without assistance of the individual database providers in the form of machine processable schemas, ontologies, terminologies, and accompanying query APIs, query languages, and standardized data exchange formats (e.g., XML) and the associated data definitions. *We need data management systems better suited to bioinformatics applications. Such systems will not appear spontaneously. The research and development for such bioinformatics data management technology will require federal funding of targeted interdisciplinary research.*

We anticipate that bioinformatics applications will be one of the major drivers for innovation in data management technology over the next decade. The resulting data management technologies should greatly facilitate the development of bioinformatics applications, and hence the conduct of biological and biomedical research over the next quarter century. To read the full workshop report, please go to the workshop website: *http://www.lbl.gov/~olken/wdmbio/*

—F. Olken
Lawrence Berkeley National Laboratory
Berkeley, California

—H.V. Jagadish
University of Michigan
Ann Arbor, Michigan